Abstract

RGB-IR Cross Modality Person Re-Identification (reID) comes under the field of computer vision and the problem has the requirement of matching a person across multiple camera views either from overlapping/non-overlapping cameras of RGB or Infrared (IR) modalities. Adding to the distribution discrepancy between the different modalities of images, nuisance factors such as background clutter, misalignment errors elevate the difficulty of the task. In this report, we propose two solutions methods - **A**ttribute **B**ased **R**epresentation **L**earning (ABRL) and **D**omain **I**nvariant **R**epresentation **L**earning (DIRL) based on novel loss functions (attribute classification loss and domain classification respectively) along with re-ranking for the task of RGB-IR cross modality person reID. In addition, we have annotated common pedestrian attribute information for two large-scale RGB-IR reID datasets. Extensive ablations studies and evaluation results shows the effectiveness of our proposed solutions. Especially, our proposed method, DIRL enhances Rank-1 accuracy by 21.3% on SYSU-MM01 dataset and ABRL enhances Rank-1 accuracy by 18% on RegDB dataset compared to state-of-the-arts.

Contents

Chapter 1

Introduction

Person re-identification (termed as person reID) comes under the domain of computer vision and the problem requires to match a person across multiple camera sources either from overlapping/non-overlapping cameras (Refer Fig. 1.1). Due to wide applicability of this task in the domains such as Surveillance, Human-computer interaction, Behavioral analysis and Home-automation, it has attained great research interest recently both from academia and industry.

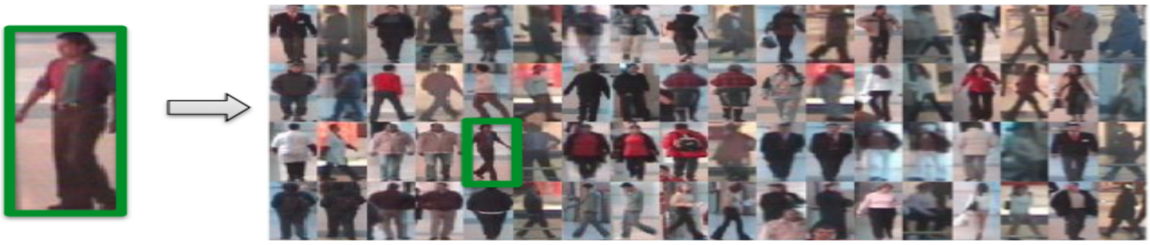


Figure 1.1: Depiction of person reID task. The person image on left-side is the probe instance and the person images after the arrow mark are gallery instances. Person reID task is to match query instance with closest gallery instances and produce a ranking of gallery instances based on its similarity with query instance. Green bounding boxes show that the images represent the same person.

Person re-identification is extensively analysed in RGB spectrum, where the images are taken by RGB cameras. However, RGB images are not helpful to re-identify a person under low light scenarios such as night-time. To provide round-the-clock functionality and security, the surveillance systems typically make use of IR cameras, as they are superior during poor illumination conditions and less dependent on the lighting. The problem of matching a person's image across multiple modalities of images such as RGB and Infrared(IR) is a case of cross modality person reID. The huge variation between the images of RGB and IR modalities makes RGB-IR cross modality reID, a very challenging problem. In addition, there exists several challenges such as pose variations, background clutter which further exacerbate the task of RGB-IR person reID. A common solution method for person reID is to learn feature descriptors of the images and compare the descriptors based on a distance metric to match persons. Convolutional neural networks(CNNs) have become a prevalent choice for learning features of persons from images in cross modality person reID due to their geometrical invariant properties when compared to traditional methods[1, 2].

Attributes describe details about a person which includes gender, type of clothes, hats, shoes, etc. Deep learning models which use attribute information for matching persons better have been proposed for single modal reID [3, 4]. In a similar way, we explore the use of attribute information for the task of RGB-IR cross modality person reID.

Domain adaptation is the task of learning domain independent features for two or more domains as input. Cross modality person reID can also be considered under the field of domain adaptation. We also explore the direction of using domain adaption techniques for RGB-IR person image matching.

Re-ranking is a well known post-processing procedure in person re-identification[20, 21] and other similar ranking tasks[5]. Typical re-ranking solutions in person reID make use of neighbourhood information to improve the initial ranking [6, 7]. Re-ranking has produced notable improvements in RGB-RGB person reID as a post-processing step. However, re-ranking is not well applied in the task of cross modality person reID. In this work, we apply re-ranking based on k-reciprocal neighbours[6] as post-processing step and analyze the performance improvements.

In this report we propose two solution methods for the task of RGB-IR person reID 1) using attributes information, 2) using domain discriminative networks [8]. We also apply re-ranking as a post-processing procedure for both the proposed solutions methods.

1.1 Motivation

My motivation for this project stems from curiosity of how learning models work and the widespread applications of reID in real world. Person reID has profound implications in security and surveillance; even a slight improvement in performance can contribute a great deal in the field of surveillance. My interests in reID peaked after working in the related task of disguised face recognition as part of my summer internship. Another aspect that draws me to this project is the satisfaction of engaging in it. Through this project, I hope to get a better understanding of deep learning models and use that knowledge in the near future.

1.2 Problem Definition

RGB-IR Cross Modality Person Re-identification is the task of matching a person across multiple camera views either from overlapping/non-overlapping cameras of RGB or Infrared (IR) modalities.

In other words, given a query image of a person the task is to retrieve images of the same person from a large gallery set which contains images from various camera views. The gallery images are to be ranked based on their similarity with query image and images can be of either RGB modality or IR modality.

1.3 Contributions

We contribute the following through our work,

1. A list of annotated common person features for the person images in SYSU-MM01 dataset and the RegDB dataset.
2. Two objective functions - 'Attribute classification loss' and 'Domain classification loss' to improve the performance in RGB-IR cross modality person re-identification scenario.
3. Ablation studies of the loss functions and model architecture to evaluate the effectiveness of our proposed methodology.

Chapter 2

Related Literature

Image-based person re-identification (reID) has seen significant improvements with the progress of deep learning. Majority of reID research in the past few years uses deep CNN models as they are robust to pose and viewpoint changes in visible-visible matching i.e, single modal reID. However, there are only few studies on visible-infrared (RGB-IR) matching i.e, cross modality reID. We discuss the related works for RGB-IR reID and touch upon attribute based reID, domain adaptation, re-ranking in this section.

2.1 RGB-IR person reID

In some of the earlier works of RGB-IR cross-modal reID, Ancong Wu *et al.*[1] introduced a deep zero-padded single stream network and showed its performance is superior than that of feature extraction methods such as HOG and LOMO [9, 2]. Mang Ye *et al.*[10] used a two stage framework for the task of RGB-IR matching. cmGAN[11] uses a generative adversarial network for learning common representations of RGB and IR images. Dual-level Discrepancy Reduction Learning (D^2RL) method manages the modality discrepancy and appearance discrepancy in RGB-IR modals separately and proves to be more effective as shown by Zhixiang Wang *et al.*[12]. Zhao *et al.*[13] used a single stream model architecture for images of both modalities and proposed a hard pentaplet loss for reducing modality discrepancy. The state of the art[14] method - AlignGAN uses deep neural architectures for joint feature and pixel alignment (Fig. 2.1).

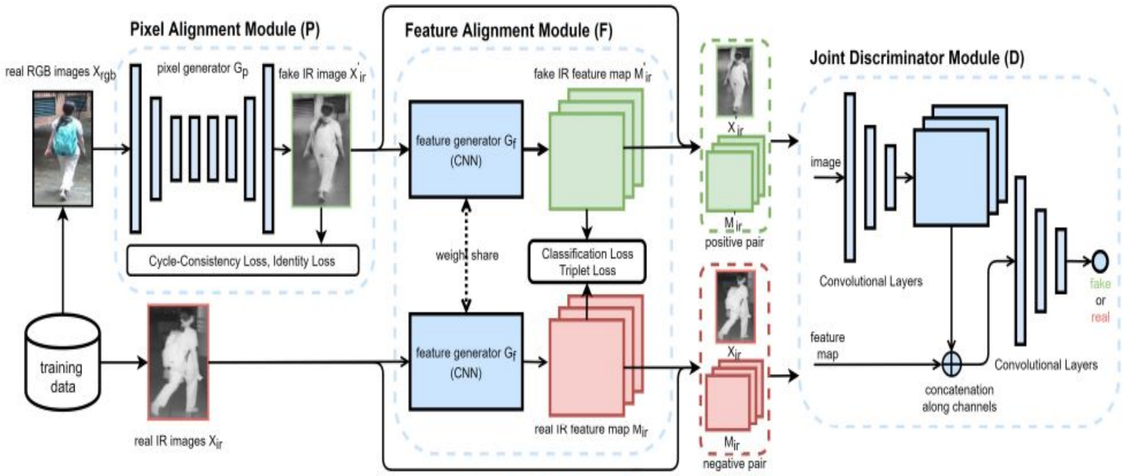

Pixel Alignment Module (P)
Feature Alignment Module (F)
Joint Discriminator Module (D)
real RGB images X_{rgb}
pixel generator G_p
fake IR image X'_{ir}
Cycle-Consistency Loss, Identity Loss
training data
real IR images X_{ir}
feature generator G_f (CNN)
weight share
feature generator G_f (CNN)
fake IR feature map M'_{ir}
Classification Loss
Triplet Loss
real IR feature map M_{ir}
X_{ir}
M'_{ir}
positive pair
X_{ir}
M_{ir}
negative pair
image
Convolutional Layers
feature map
concatenation along channels
Convolutional Layers
fake
or
real

FIGURE 2.1: Pipeline of AlignGAN framework which uses pixel and feature alignment modules for RGB-IR person reID

2.2 Attribute based person reID

Attributes describe details about a person which includes gender, type of clothes, hats, shoes, etc. Attributes have been investigated for deep models of person re-identification [15, 16, 17, 18]. Chi Su *et al.*[16] proposed a cross-dataset feature learning system on the principle that same person should have same attributes. In [17], a siamese network is proposed for joint attribute recognition and identity verification, which treats the verification loss as the heart of multi-task. In this report, we consider attribute based discriminative learning for the task of RGB-IR person reID through a simple loss function and would build on this in future.

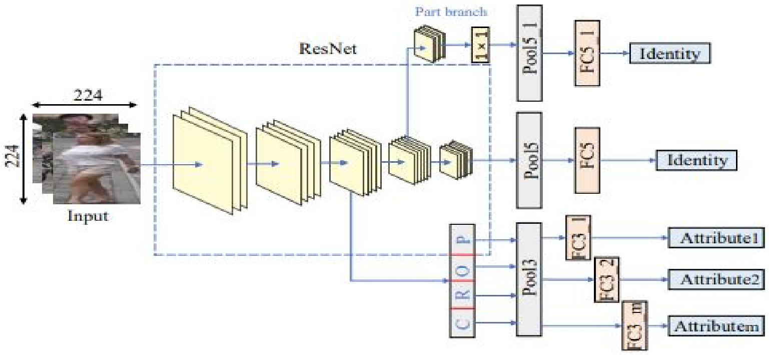


FIGURE 2.2: Illustration of an attribute based CNN framework for person reID

2.3 Domain adaptation and person reID

Cross modality person reID can be considered as unsupervised domain adaptation because for a given RGB image there is no corresponding IR image available during training and vice versa. Several prior research related to unsupervised domain adaptation across datasets have been done in the context of person reID [19, 20]. Domain confusion loss was introduced in [21, 22] for aiding the model to learn domain-invariant features (Refer Fig. 2.3). In this report, we consider learning domain-invariant features using gradient reversal method [8] as another solution direction.

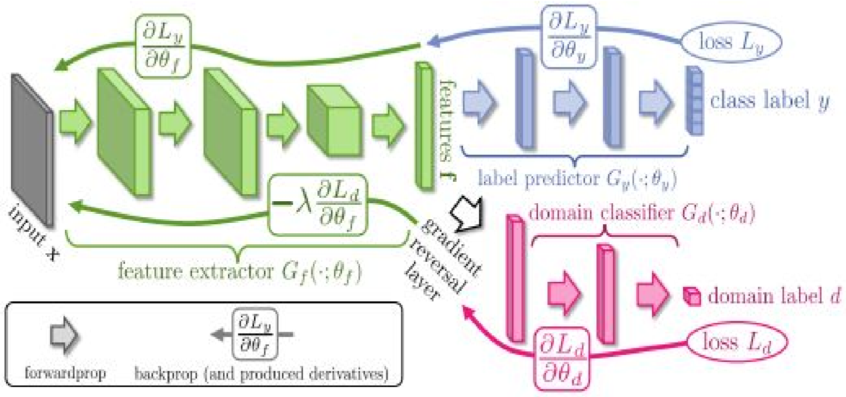

$\frac{\partial L_y}{\partial \theta_f}$
$\frac{\partial L_y}{\partial \theta_y}$
loss L_y
features f
class label y
label predictor $G_y(\cdot;\theta_y)$
$-\lambda\frac{\partial L_d}{\partial \theta_f}$
domain classifier $G_d(\cdot;\theta_d)$
input x
gradient reversal layer
feature extractor $G_f(\cdot;\theta_f)$
domain label d
loss L_d
$\frac{\partial L_d}{\partial \theta_d}$
forwardprop
backprop (and produced derivatives)

FIGURE 2.3: Illustration of Domain Adversarial Network and domain confusion loss.

2.4 Re-ranking

Re-ranking is a well known post-processing procedure in person re-identification[23, 24] and other similar ranking tasks[5]. In this report, we explore the effectiveness of re-ranking in RGB-IR person reID. We employ re-ranking based on k-reciprocal encoding [6] which enhances initial ranking using neighbourhood information as a post-processing step in our methods. Any re-ranking strategy requires a matching between probe or query set of images/features and gallery set of images/features to produce a ranking of gallery instances for each probe instance. The paper[6] on k-reciprocal encoding introduces a re-ranking strategy as per the followings steps,

- *'K-reciprocal nearest neighbour encoding'* and *'local query expansion'*

  K-reciprocal nearest neighbour set for a probe instance (say p_i) contains those gallery instances which are in the k-nearest neighbourhood of p_i and also contain p_i in their k-nearest neighbourhood. K-reciprocal nearest neighbours are encoded in vector form and local query expansion is used to make the vector more informative.

- *'Weighted distance computation'*

  Jaccard distance ($D_{jaccard}$) on k-reciprocal feature is used to measure the distance between gallery and probe instances. Since, the original distance matrix ($D_{original}$) also captures valid information, a weighted aggregate of $D_{original}$ and $D_{jaccard}$ is used to obtain the final score matrix.

Chapter 3

Methodology

3.1 Attribute based representation learning

Deep learning models which use attribute information have two-fold benefits. Firstly, using attribute labels in training process helps in increasing the discriminative ability of the network used. Secondly, attributes can help in expediting the retrieval process by pruning gallery instances which have different attributes when compared to the query instance. This made us to consider using attributes to aid the feature learning process in RGB-IR reID. We have manually annotated common pedestrian attributes such as gender, type of clothes, backpack, etc for the RGB-IR datasets. We propose a loss function for learning attribute based features for the task of RGB-IR reID and we name this method as **A**ttribute **B**ased **R**epresentation **L**earning (ABRL).

3.2 Domain-Invariant representation learning

Adversarial networks are applied in several tasks which require learning domain invariant features[25, 26]. The task of RGB-IR reID can be formulated as learning RGB/IR invariant features and the method of adversarial feature learning can be applied here. Ganin Yaroslav *et al.*[8] introduced gradient reversal through back propagation for adversarial training. Our solution method is to learn domain-invariant features through gradient reversal networks and we name our model as **D**omain **I**nvariant **R**epresentation **L**earning (DIRL).

3.3 Pipeline

Our proposed pipeline (Refer to Figure 3.1) consists of four stages as follows,

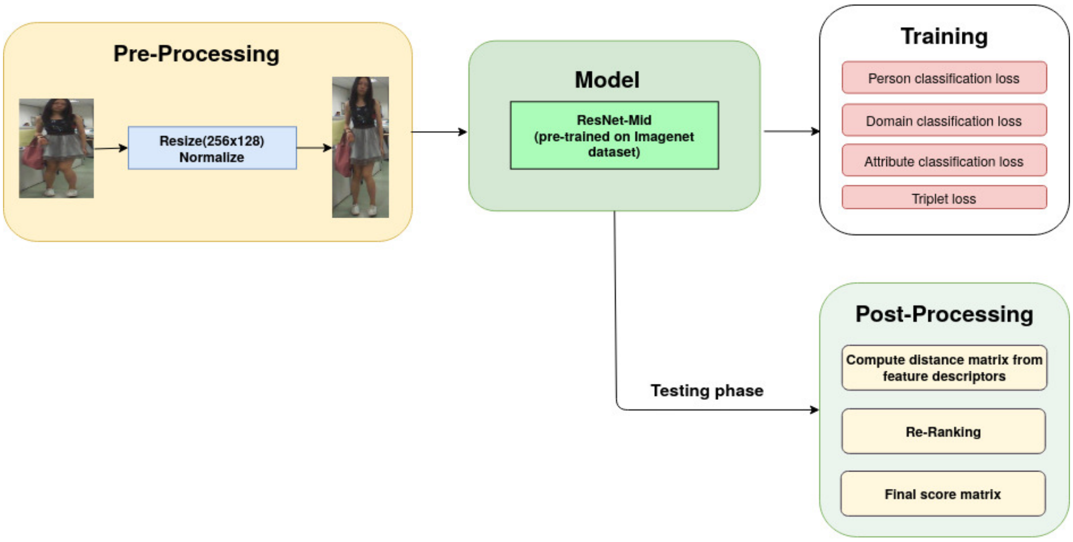

Pre-Processing
Resize(256x128)
Normalize
Model
ResNet-Mid
(pre-trained on Imagenet dataset)
Training
Person classification loss
Domain classification loss
Attribute classification loss
Triplet loss
Testing phase
Post-Processing
Compute distance matrix from feature descriptors
Re-Ranking
Final score matrix

FIGURE 3.1: The figure shows our proposed pipeline. The pre-processing phase involves resizing the image & normalizing the image, Resnet-Mid[27] is used as backbone network, training phase uses four loss functions (Refer section 3.3.3). Testing phase involves normalizing the feature descriptors from the model and applying re-ranking[6] procedure.

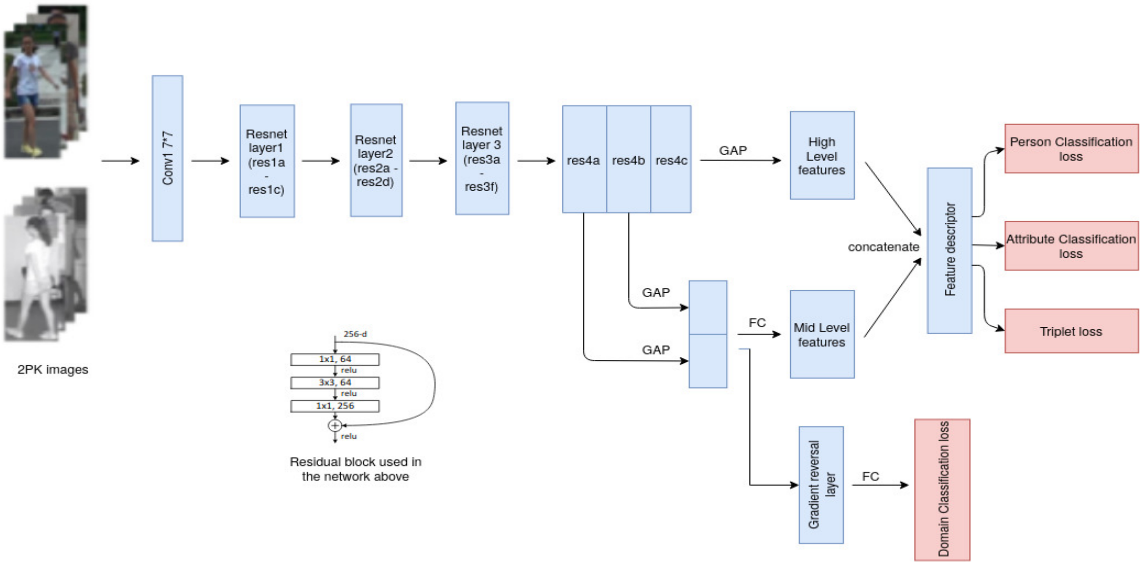

2PK images
Conv1 7*7
Resnet layer1 (res1a - res1c)
Resnet layer2 (res2a - res2d)
Resnet layer 3 (res3a - res3f)
res4a
res4b
res4c
GAP
High Level features
GAP
GAP
FC
Mid Level features
concatenate
Feature descriptor
Person Classification loss
Attribute Classification loss
Triplet loss
Gradient reversal layer
FC
Domain Classification loss
256-d
1x1, 64
relu
3x3, 64
relu
1x1, 256
relu
Residual block used in the network above

FIGURE 3.2: Depiction of our model architecture with objective functions. Here, 2PK refers to number of training samples in a batch where, P = count of distinct persons in a training batch & K = count of images per modality for each distinct person in training batch.

3.3.1 Pre-processing

Pre-processing phase consists of the following steps,

1. For maintaining proper human proportions, the input person images are resized to a height of 256 and a width of 128.
2. The images are then normalized with a mean of '$[0.485, 0.456, 0.406]$' and standard deviation of '$[0.229, 0.224, 0.225]$' across R,G,B channels.

3.3.2 Model architecture

Our model architecture uses ResnetMid[27] as the backbone network (Refer Fig 3.2) which is trained along with a combination of various loss functions (Section 3.3.3) to tackle the task of RGB-IR cross modality person reID. Based on the objective functions used the methods are classified into ABRL and DIRL.

3.3.3 Objective Functions

Our solution methods use a combination of the following objective functions:

3.3.3.1 Person classification loss

The standard cross-entropy loss with label smoothing[28] is employed to make the model learn discriminative features for identifying a person. For an input image I_i, the target identity of the person is encoded in one-hot format as t_i and the softmax output of the person identity p_i is obtained from the model. Then, the person classification loss is calculated as follows,

$$\begin{aligned} \hat{t}_i &= (1-\epsilon)\cdot t_i + \frac{\epsilon}{M} \\ L_{id} &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\hat{t}_{ij}\log p_{ij} \end{aligned} \tag{3.1}$$

Here, N = batch size, M = number of distinct persons in training dataset, ϵ = smoothing parameter (Here we use $\epsilon = 0.1$). The label smoothing of t_i acts as a regularizer and helps in better generalization.

3.3.3.2 Triplet loss

We employ Triplet loss in addition to person classification loss for learning discriminative features of the person identities. For each image embedding, I_i obtained from the model in a batch of training samples, hard positive image embedding I_{i+} are defined as those embedding that are of the same class as that of I_i & are farthest from I_i based on a distance metric. We similarly define hard negative instances, I_{i-} as the embeddings which are closest to I_i and are of different class compared to I_i in the training batch. These instances are mined[29] for each image in the training batch. We define triplet loss as follows,

$$L_{trip} = \frac{1}{N}\sum_{i=1}^{N} max(0, d(I_i, I_{i+}) - d(I_i, I_{i-}) + m) \tag{3.2}$$

Here N = batch size, $d(i, j)$ = distance between the image embeddings i & j (Here, we use Euclidean distance), m = margin parameter.

3.3.3.3 Domain classification loss

To make the model invariant to the modality of person images, we make use of cross entropy loss with label smoothing[28] after *gradient reversal layer.* The target identity for each image, I_i is their modality (RGB image or IR image) and it is encoded in one-hot format as t_i. Given the softmax output of the modality p_i from the model after the *gradient reversal layer*, we define the domain classification loss as,

$$\begin{aligned} \hat{t}_i &= (1-\epsilon)\cdot t_i + \frac{\epsilon}{M} \\ L_{domain} &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} \hat{t}_{ij}\log p_{ij} \end{aligned} \tag{3.3}$$

Here, ϵ = smoothing parameter (we use $\epsilon = 0.1$), N = batch size, M = Number of modalities (Here, $M = 2$ as there are only RGB and IR modalities).

3.3.3.4 Attribute classification loss

To make the model learn pedestrian attribute based features, we employ a cross-entropy classification loss with label smoothing for each attribute field. Given a person's image I_i, the target attribute identity is encoded in one-hot format as t_{ij} for each attribute field j and the softmax output of attribute identity p_{ij} is obtained from the model for each attribute field j. Then, the attribute classification loss is defined as follows,

$$\begin{gathered}\hat{t}_{ij} = (1 - \epsilon) \cdot t_{ij} + \frac{\epsilon}{L_j} \\ L_{attribute} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{L_j} \hat{t}_{ijk} \log p_{ijk}\end{gathered} \tag{3.4}$$

Here, ϵ = smoothing parameter (we use $\epsilon = 0.1$), N = batch size, M = number of attribute fields, L_j = number of options for attribute field j.

3.3.3.5 Overall Objective function

The overall loss functions for Attribute Based Representation Learning (ABRL) and Domain Invariant Representation Learning (DIRL) are given by

$$L_{ABRL} = \gamma_1 L_{id} + \gamma_2 L_{trip} + \gamma_3 L_{attribute} \tag{3.5}$$

$$L_{DIRL} = \alpha_1 L_{id} + \alpha_2 L_{trip} + \alpha_3 L_{domain} \tag{3.6}$$

The ratios $\gamma_1 = 10, \gamma_2 = 1, \gamma_3 = 1, \alpha_1 = 10, \alpha_2 = 1, \alpha_3 = 1$ for training on SYSU dataset are selected using validation set. The ratios $\gamma_1 = 1, \gamma_2 = 10, \gamma_3 = 1, \alpha_1 = 1, \alpha_2 = 10, \alpha_3 = 1$ are used for training on RegDB dataset.

3.3.4 Post-processing

Post-processing phase (See Fig. 3.1) has the following steps,

1. We use Euclidean distance to find the distance between query & gallery features obtained from the model and form the distance matrix.
2. We apply re-ranking[6] on the distance matrix to obtain the processed distance matrix which we use for evaluation.

Chapter 4

Experiments

4.1 Datasets, Implementation, Evaluation protocol

4.1.1 Datasets

Two RGB-IR reID datasets have been identified for evaluation of learning models in the task of RGB-IR cross modality person reID.

1. **SYSU-MM01 RGB-IR dataset**[1] is a dataset of 303,420 person images acquired by 6 cameras (four RGB and two IR). The dataset has 491 persons with the decomposition of training set containing 296 persons, validation set of 99 persons and testing set containing 96 persons. Three cameras are located in outdoor settings and the rest in indoor settings. Each person has images from at least two cameras.
2. **RegDB RGB-IR dataset**[30] is dataset of 412 persons where the images are acquired by two cameras. Each person has 20 images in total which are acquired by RGB and IR cameras.

4.1.2 Implementation, Training

Our models are implemented using Pytorch framework with GPU acceleration. We use ADAM[31] optimizer with $5e^{-4}$ as weight decay and a learning rate of $3e^{-4}$ to optimize the model for 50 epochs. Refer to section 3.3.1 for pre-processing steps which include resizing and normalization. The training samples are randomly flipped for data augmentation and a training batch size of 64 is utilized. For each batch, we randomly sample 16 identities with 2 images from RGB modality and 2 images from IR modality. We use a margin of 0.3 for L_{trip} and we decay the learning rate after 10 epochs at a decay rate of 0.1. The output feature of an input image is a vector of dimensions 3048.

4.1.3 Evaluation protocol

We utilize the following evaluation metrics, 1) mean average precision (mAP) and 2) Rank-N accuracy or the Cumulative Matching Characteristic (CMC) with $N \in \{1, 10, 20\}$ for evaluating models in the task of RGB-IR cross modality person reID. SYSU-MM01 has two evaluation protocols based on location ('indoor search' and 'all search') and for each of them there are 'multi-shot' and 'single-shot' settings, giving rise to an overall of four evaluation protocols. The gallery set contains RGB images and probe set contains IR images for SYSU dataset. For regDB dataset, there are two evaluation protocol based on the modality of probe/gallery set and not based on shot-setting/location. The evaluation modes are *thermal2visible* and *visible2thermal*. *Thermal2visible* means that the probe set consists only of thermal images, gallery set consists only of visible images and *visible2thermal* is defined similarly. Note that RegDB follows a different naming convention, i.e, visible for RGB modality and thermal for IR modality. For consistency, the results are averaged over 10 times using random split of the dataset.

4.2 Attribute fields

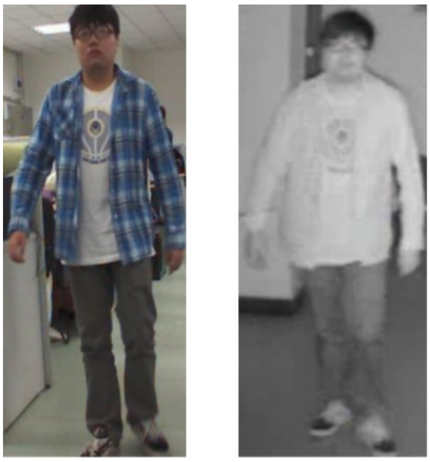
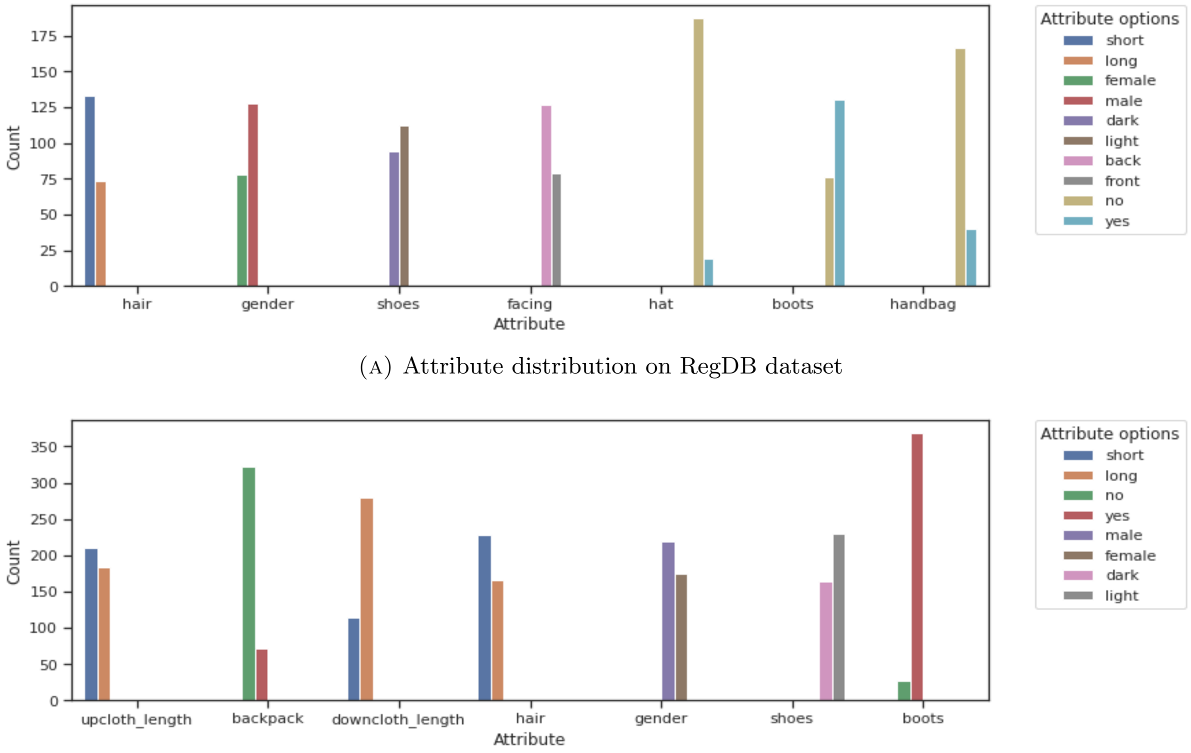


FIGURE 4.1: Attributes of the person based on the given images : male, adult, pants, etc. Sample images are from SYSU dataset.

Attribute field	Options
gender	male, female
age	young, teen, adult, old
backpack	yes, no
boots	yes, no
clothes	dress, pants
downcloth length	long, short
upcloth length	long, short
facing	front, back
hair	long, short
hat	yes, no
handbag	yes, no
shoes	dark, light

TABLE 4.1: Attribute fields and their corresponding options which have been manually annotated by us.

For manually annotating common pedestrian attribute fields, we choose fields (refer Table 4.1) which can be inferred from both the modalities. We avoid using colour attributes as they cannot be inferred from IR/Thermal images.



(A) Attribute distribution on RegDB dataset

(B) Attribute distribution on SYSU dataset

FIGURE 4.2: Some of the attribute's distribution on training samples of RegDB dataset(Fig. 4.2a) and SYSU dataset (Fig. 4.2b).

4.3 Analysis of objective functions

We perform ablation studies on the losses - attribute loss ($L_{attribute}$) and domain invariance loss (L_{domain}) to find their effectiveness. We illustrate the results for SYSU dataset on the hard evaluation case of 'all-search' and 'single-shot' setting and for RegDB dataset on the evaluation case of 'visible to thermal' or 'RGB to IR' setting. In this study, the weightage for the person identity loss (L_{id}) and triplet loss (L_{trip}) are kept constant ($\gamma_1 = 10, \gamma_2 = 1$ for SYSU dataset and $\gamma_1 = 1, \gamma_2 = 10$ for RegDB dataset). We experiment with performance improvements of adding the proposed losses. From Tab. 4.2, we can see that both the losses improve the performance of the model. In SYSU dataset L_{domain} loss improves the performance by the most, while in RegDB dataset both the losses perform similarly.

Losses		SYSU dataset				RegDB dataset			
		All-search single-shot				Visible to Thermal			
L_{domain}	$L_{attribute}$	R1	R10	R20	maP	R1	R10	R20	maP
		43.7	82.1	92.1	44.1	64.3	84.8	91.4	61.4
	✓	44.8	83.7	92.2	44.7	69.2	88.7	93.5	64.8
✓		48.1	86.8	94.6	46.8	69.9	85.3	90.8	64.6

Table 4.2: Results of resnet-mid model trained with various configurations of the proposed objective functions: Attribute loss ($L_{attribute}$), Domain loss (L_{domain}) on SYSU and RegDB dataset.

4.4 Architecture of domain classifier

In this section, we experiment with the architecture of the domain classifier which is succeeding the gradient reversal layer. We consider three architectures with increasing number of layers and tabulate our results (Refer Tab 4.3). We define $Block(in\_dims,\ out\_dims)$ which will be part of our layers as follows,

$$\begin{aligned} Block(in\_dims,\ out\_dims) = [\ & Linear(in\_dims,\ out\_dims), \\ & BatchNorm(out\_dims), \\ & Relu(out\_dims)\ ] \end{aligned}$$

Architecture	SYSU dataset All-search single-shot				RegDB dataset Visible to Thermal			
	R1	R10	R20	maP	R1	R10	R20	maP
Linear(4096, 2)	42.7	83.7	92.4	43.7	59.1	78.4	87.8	58.5

Block(4096, 2048)
Linear(2048, 2)
	48.1	86.8	94.6	46.8	69.9	85.3	90.8	64.6

Block(4096, 2048)
Block(2048, 1024)
Linear(1024, 2)
	44.1	85.2	93.5	43.2	67.0	85.9	91.2	63.1

Table 4.3: Comparison of different configurations of domain classifiers on SYSU dataset and RegDB dataset.

We observe that domain classifier with $[Block(4096, 2048),\ Linear(2048, 2)]$ gives the best performance in both the datasets.

4.5 Re-ranking hyper-parameters tuning

The Re-ranking procedure based on k-reciprocal encoding (Refer Sec. 2.4) has three hyper-parameters - *k1, k2, lambda*. *k1* controls the number of nearest neighbours being considered for a query instance, *k2* is a hyper-parameter part of local query expansion step and *lambda* controls the weight-age of original distance and Jaccard distance. SYSU dataset has separate validation dataset and hence we perform our hyper-parameter search on it. We use ResnetMid[27] trained with L_{id} and L_{trip} losses for this search and report their performance in Tab. 4.4. Based on the experiment, we choose $k1 = 22$, $k2 = 6$, $lambda = 0.3$ as hyper-parameters for all other models as it results in best performance.

Hyper-parameters			SYSU dataset			
			All-search, single-shot			
k1	k2	lambda	R1	R10	R20	maP
22	4	0.3	34.7	58.7	82.3	36.4
		0.4	34.6	58.9	82.4	36.3
	6	0.3	35.3	60.1	82.5	37.3
		0.4	35.0	60.2	82.7	37.2
24	4	0.3	34.5	58.5	82.2	36.3
		0.4	34.6	58.8	82.3	36.3
	6	0.3	35.1	60.0	82.4	37.2
		0.4	34.7	60.0	82.6	37.1

TABLE 4.4: Results of re-ranking's hyper-parameters tuning on SYSU dataset.

Chapter 5

Results

We report the performance of our methods - ABRL, DIRL on SYSU and RegDB datasets in this section and discuss their effectiveness for the problem of RGB-IR cross modality person reID.

Methods	All-search							
	Single-shot				Multi-shot			
	R1	R10	R20	mAP	R1	R10	R20	mAP
HOG [9]	2.76	18.3	32.0	4.24	3.82	22.8	37.7	2.16
LOMO [2]	3.64	23.2	37.3	4.53	4.70	28.3	43.1	2.28
Zero-padding [1]	14.8	52.2	71.4	16.0	19.2	61.4	78.5	10.9
cmGAN [12]	27.0	67.5	80.6	27.8	31.5	72.7	85.0	22.3
HPILIN [13]	41.3	84.7	94.5	42.9	47.5	88.1	95.9	36.0
AlignGAN [14]	42.4	85.0	93.7	40.7	51.5	89.4	95.7	33.9
*ABRL(ours)**	55.7	85.4	92.9	54.0	47.4	87.3	94.2	39.4
*DIRL(ours)**	**63.7**	**91.4**	**96.2**	**61.7**	**52.0**	**92.5**	**97.2**	**48.6**

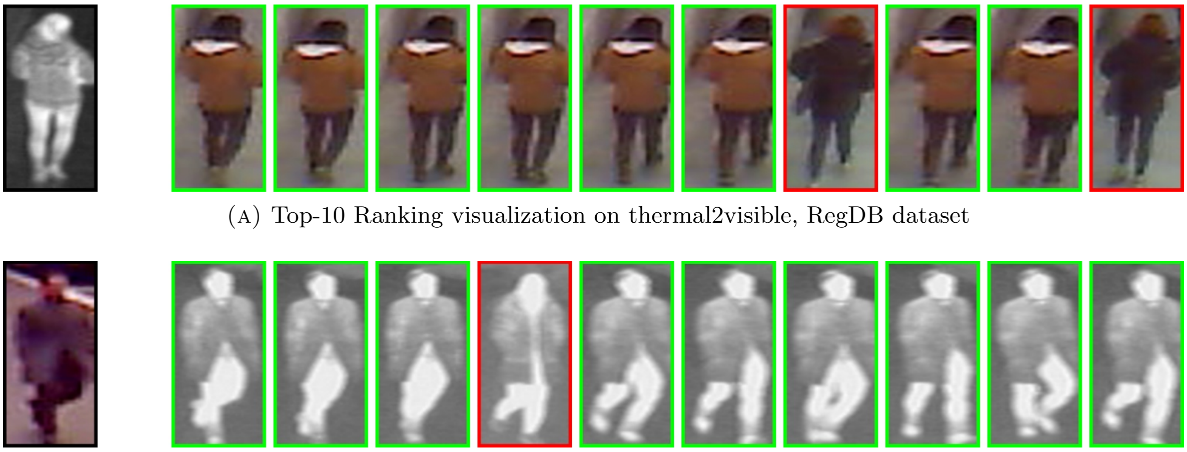
Table 5.1: Performance of various methods in the all-search setting on SYSU-MM01 dataset. The R-X denote Rank-X accuracy (%) where $X \in \{1, 10, 20\}$. The mAP corresponds to mean average precision score (%). Here, * denotes model with re-ranking procedure applied.

We observe from Table 5.1 that our methods - DIRL and ABRL surpasses the results of present methods on SYSU dataset in 'all-search' mode with 'single-shot' and 'multi-shot' setting. Our DIRL method enhances R1 accuracy and mAP at the harder case of 'Single-Shot, all-search' case in SYSU dataset by almost 20%. ABRL also improves the performance

Methods	Indoor-search							
	Single-shot				Multi-shot			
	R1	R10	R20	mAP	R1	R10	R20	mAP
HOG [9]	3.22	24.7	44.6	7.25	4.75	29.1	49.4	3.51
LOMO [2]	5.75	34.4	54.9	10.2	7.36	40.4	60.4	5.64
Zero-padding [1]	20.6	68.4	85.8	27.0	24.5	75.9	91.4	18.7
cmGAN [12]	31.7	77.2	89.2	42.2	37.0	80.9	92.3	32.8
HPILIN [13]	45.7	91.8	**98.4**	56.5	53.0	93.7	**98.9**	47.4
AlignGAN [14]	45.9	87.6	94.4	54.3	57.0	92.7	97.4	45.3
ABRL(ours)*	58.2	91.2	96.6	65.5	55.9	92.1	97.1	50.5
DIRL(ours)*	**64.9**	**93.2**	97.2	**71.0**	**67.9**	**94.7**	98.7	**61.5**

TABLE 5.2: Performance of various methods in the indoor-search setting on SYSU-MM01 dataset. The R-X denote Rank-X accuracy (%) where $X \in \{1, 10, 20\}$. The mAP corresponds to mean average precision score (%). Here, * denotes model with re-ranking procedure applied.

by a significant amount in 'all-search' setting. From Table 5.2, we observe that our methods improve significantly on Rank-N ($N \in \{1, 10\}$) and mAP evaluation metrics in 'indoor-search' protocols of SYSU dataset, while performing on par with the state-of-the-arts on Rank-20 metric.



(A) Top-10 Ranking visualization on thermal2visible, RegDB dataset

(B) Top-10 Ranking visualization on visible2thermal, RegDB dataset

FIGURE 5.1: DIRL's top-10 ranking visualization of gallery instances for the given query image on RegDB dataset in thermal2visible protocol (Fig. 5.1a) and visible2thermal protocol (Fig. 5.1b). The left image is the query instance while the right images are the gallery instances. A green bounding box implies that the person in gallery instance and query instance match while red means they don't match.

From Fig. 5.1, we see that our method DIRL is able to rank many true positives in the top-10.

However, some false negatives also come up occasionally in the top-10 ranking. We believe this is due to the background clutter in the IR and RGB images that elevate the difficulty of the task.

Methods	thermal2visible		visible2thermal	
	Rank-1	mAP	Rank-1	mAP
HOG [9]	13.4	10.3	13.9	11.1
Zero padding [1]	16.7	17.9	17.8	18.9
TONE [10]	21.7	22.3	24.4	20.1
BDTR [32]	32.8	31.2	33.5	31.9
AlignGAN [14]	56.3	53.4	57.9	53.6
*ABRL(ours)**	74.3	77.8	**75.3**	**78.3**
*DIRL(ours)**	**74.9**	**77.9**	73.7	76.7

Table 5.3: Comparison of various methods on RegDB dataset. The mAP corresponds to mean average precision score (%). Here, * denotes model with re-ranking procedure applied.

We observe from Table 5.3 that a given method performs similarly with respect to the two protocols of RegDB dataset (thermal2visible and visible2thermal). Our method, ABRL seems to perform better than DIRL in the protocol of visible2thermal while DIRL performs better than ABRL in the protocol of thermal2visible. Both the methods enhance the results on RegDB dataset by almost 20% when compared to other methods. Thus, showing the effectiveness of our DIRL and ABRL methods.

Chapter 6

Conclusions and Future work

Our work is intended to improve person identification performance for surveillance in night time by the usage of infrared cameras. The huge variation between the images of two modalities (RGB and IR) makes RGB-IR cross modality reID, a very challenging problem. In this project, we discuss two solutions based on deep neural network structures for the cross domain task and pave way for surveillance and security companies to improve their products by providing a robust person re-identification system. We observe that attribute information can be used to make a model more robust and we plan to explore other techniques in the available literature[33, 18] for using attribute information in aiding the training process in future.

Bibliography

[1] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, "RGB-Infrared cross-modality person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5380–5389.

[2] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2197–2206.

[3] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, vol. 95, p. 151–161, Nov 2019. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2019.06.006

[4] A. Schumann and R. Stiefelhagen, "Person re-identification by deep learning attribute-complementary information," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 20–28.

[5] A. Subramaniam, A. Narayanan Sridhar, and A. Mittal, "Feature ensemble networks with re-ranking for recognizing disguised faces in the wild," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019, pp. 532–541.

[6] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1318–1327.

[7] M. Ye, C. Liang, Y. Yu, Z. Wang, Q. Leng, C. Xiao, J. Chen, and R. Hu, "Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2553–2566, 2016.

[8] Y. Ganin and V. S. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32nd International Conference on Machine Learning, ICML*

2015, Lille, France, 6-11 July 2015, ser. JMLR Workshop and Conference Proceedings, vol. 37. JMLR.org, 2015, pp. 1180–1189. [Online]. Available: http://proceedings.mlr.press/v37/ganin15.html

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1, 2005, pp. 886–893.

[10] M. Ye, X. Lan, J. Li, and P. C. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 7501–7508.

[11] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, "Cross-modality person re-identification with generative adversarial training," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018, pp. 677–683.

[12] Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, "Learning to reduce dual-level discrepancy for infrared-visible person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 618–626.

[13] Y.-B. Zhao, J.-W. Lin, Q. Xuan, and X. Xi, "HPILN: a feature learning framework for cross-modality person re-identification," *IET Image Processing*, vol. 13, no. 14, p. 2897–2904, 12 2019. [Online]. Available: http://dx.doi.org/10.1049/iet-ipr.2019.0699

[14] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "RGB-infrared cross-modality person re-identification via joint pixel and feature alignment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3623–3632.

[15] A. Li, L. Liu, and S. Yan, *Person Re-identification by Attribute-Assisted Clothes Appearance.* Springer, 2014, pp. 119–138.

[16] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *European conference on computer vision.* Springer, 2016, pp. 475–491.

[17] N. McLaughlin, J. M. del Rincon, and P. C. Miller, "Person reidentification using deep convnets with multitask learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 525–539, 2016.

[18] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, vol. 95, pp. 151–161, 2019.

[19] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *European conference on computer vision.* Springer, 2010, pp. 213–226.

[20] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2066–2073.

[21] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[22] H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, and M. Marchand, "Domain-Adversarial Neural Networks," *arXiv e-prints*, p. arXiv:1412.4446, Dec. 2014.

[23] G. Wang, J. Lai, P. Huang, and X. Xie, "Spatial-temporal person re-identification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, p. 8933–8940, Jul 2019.

[24] P. Pathak, A. Erfan Eshratifar, and M. Gormish, "Video Person Re-ID: Fantastic Techniques and Where to Find Them," *arXiv e-prints*, p. arXiv:1912.05295, Nov. 2019.

[25] Y. Zhang, Y. Zhang, Y. Wang, and Q. Tian, "Domain-Invariant Adversarial Learning for Unsupervised Domain Adaption," *arXiv e-prints*, p. arXiv:1811.12751, Nov. 2018.

[26] Z. Meng, J. Li, and Y. Gong, "Attentive adversarial learning for domain-invariant training," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6740–6744.

[27] Q. Yu, X. Chang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "The Devil is in the Middle: Exploiting Mid-level Representations for Cross-Domain Instance Matching," *arXiv e-prints*, p. arXiv:1711.08106, Nov. 2017.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[29] A. Hermans, L. Beyer, and B. Leibe, "In Defense of the Triplet Loss for Person Re-Identification," *arXiv e-prints*, p. arXiv:1703.07737, Mar. 2017.

[30] D. Nguyen, H. Hong, K. Kim, and K. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors*, vol. 17, no. 3, p. 605, 2017.

[31] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[32] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, “Visible thermal person re-identification via dual-constrained top-ranking,” in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018, pp. 1092–1099.

[33] G. Zhang and J. Xu, “Person re-identification by mid-level attribute and part-based identity learning,” in *Asian Conference on Machine Learning*, 2018, pp. 220–231.